Convergence Analysis and Adaptively Weighted Regularization for Multiple Kernel Learning (Fundamental Technologies for the Next-Generation Computational Science)

AUTHOR(S):

Suzuki, Taiji

数理解析研究所講究録
第1848巻 2013年 61-76

Convergence Analysis and Adaptively Weighted Regularization for Multiple Kernel Learning

東京大学大学院 数理情報学専攻　鈴木 大慈 (Taiji Suzuki)
Graduate School of Information Science and Technology, University of Tokyo

Abstract

In this article, we survey recent developments of generalization error analysis of Multiple Kernel Learning (MKL) and a refined method based on the theoretical developments. The main target in this article is dense type regularizations including ℓ_p-MKL that imposes ℓ_p-mixed-norm regularization instead of ℓ_1-mixed-norm regularization. According to the recent numerical experiments, the sparse regularization does not necessarily show a good performance compared with dense type regularizations. Motivated by this fact, a general theoretical tool was recently established to derive fast learning rates that is applicable to arbitrary mixed-norm-type regularizations in a unifying manner. As a by-product, the general result gives a fast learning rate of ℓ_p-MKL that is tightest among existing bounds. The learning rate achieves the minimax lower bound. As a consequence, when the complexities of candidate reproducing kernel Hilbert spaces are inhomogeneous, it is shown that dense type regularization shows better learning rate compared with sparse ℓ_1 regularization. Moreover, on the basis of the theoretical analysis, a new method of MKL that utilizes an adaptively weighted regularization has been proposed. The method controls strength of penalty for each kernel depending on its importance so that important components are amplified and unimportant components are shrinked.

1 Introduction

Multiple Kernel Learning (MKL) proposed by [20] is one of the most promising methods that adaptively select the kernel function in supervised kernel learning. A kernel method is widely used and several studies have supported its usefulness [25]. However the performance of kernel methods critically relies on the choice of the kernel function. Many methods have been proposed to deal with the issue of kernel selection. [23] studied hyperkrenels as a kernel of kernel functions. [2] considered DC programming approach to learn a mixture of kernels with continuous parameters. Some studies tackled a problem to learn non-linear combination of kernels as in [4, 9, 37]. Among them, learning a linear combination of finite candidate kernels with non-negative coefficients is the most basic, fundamental and commonly used approach. The seminal work of MKL by [20] considered learning convex combination of candidate kernels. This work opened up the sequence of the MKL studies. [5, 22] showed that MKL can be reformulated as a kernel version of the group lasso [39]. This formulation gives an insight that MKL can be described as a ℓ_1-mixed-norm regularized method. As a generalization of MKL, ℓ_p-MKL that imposes ℓ_p-mixed-norm regularization has been proposed [22, 14]. ℓ_p-MKL includes the original

MKL as a special case as ℓ_1-MKL. Another direction of generalizing MKL is elasticnet-MKL [26, 34] that imposes a mixture of ℓ_1-mixed-norm and ℓ_2-mixed-norm regularizations. Recently numerical studies have shown that ℓ_p-MKL with $p > 1$ and elasticnet-MKL show better performances than ℓ_1-MKL in several situations [14, 8, 34]. An interesting perception here is that both ℓ_p-MKL and elasticnet-MKL produce denser estimator than the original ℓ_1-MKL while they show favorable performances.

In the pioneering paper of [20], a convergence rate of MKL is given as $\sqrt{\frac{M}{n}}$, where M is the number of given kernels and n is the number of samples. [27] gave improved learning bound utilizing the pseudo-dimension of the given kernel class. [38] gave a convergence bound utilizing Rademacher chaos and gave some upper bounds of the Rademacher chaos utilizing the pseudo-dimension of the kernel class. [8] presented a convergence bound for a learning method with L_2 regularization on the kernel weight. [10] gave the convergence rate of ℓ_p-MKL as $\frac{M^{1-\frac{1}{p}} \vee \sqrt{\log(M)}}{\sqrt{n}}$ for $1 \leq p \leq 2$. [15] gave a similar convergence bound with improved constants. [16] generalized this bound to a variant of the elasticnet type regularization and widened the effective range of p to all range of $p \geq 1$ while in the existing bounds $1 \leq p \leq 2$ was imposed. One concern about these bounds is that all bounds introduced above are "global" bounds in a sense that the bounds are applicable to all candidates of estimators. Consequently all convergence rate presented above are of order $1/\sqrt{n}$ with respect to the number n of samples. However, by utilizing the *localization* techniques including so-called local Rademacher complexity [6, 17] and peeling device [35], we can derive a faster learning rate. Instead of uniformly bounding all candidates of estimators, the localized inequality focuses on a particular estimator such as empirical risk minimizer, thus can gives a sharp convergence rate.

Localized bounds of MKL have been given mainly in sparse learning settings [18, 21, 19], and there are only few studies for non-sparse settings in which the sparsity of the ground truth is not assumed. [13] gave a localized convergence bound of ℓ_p-MKL. However, their bound is a little bit larger than the minimax optimal rate.

Recently [31, 30] gave a unified framework to derive fast convergence rates of MKL with various regularization types. The framework is applicable to *arbitrary* mixed-norm regularizations including ℓ_p-MKL and elasticnet-MKL. The derived learning rate utilizes the localization technique, thus is tighter than global type learning rates. Moreover the analysis deals with more general regularization than that of [13]. It is shown that the bound achieves the minimax-optimal rate. As a by-product, it gives a tighter convergence rate of ℓ_p-MKL than existing results. According to the analysis, dense type regularizations can outperforms sparse ℓ_1 regularization when the *complexities* of the RKHSs are not uniformly same. As far as the authors know, this research is the first theoretical attempt to clearly show advantage of dense type MKL.

On the basis of the theoretical analysis by [31, 30], [32] proposed a new MKL method that utilizes an adaptively tailored regularization to improve the performance. The method consists of two stages. In the first stage, it computes a rough estimator to approximate the true function. Then in the second stage, it constructs an adaptively weighted penalty based on the rough estimator obtained in the first stage, and compute an estimator using

the adaptively weighted penalty. The adaptive weight is intended to amplify important components and shrink unimportant components. The method can be seen as the MKL version of *adaptive lasso* [40], but the framework involves more general regularizations than ℓ_1-regularization.

This article gives an overview of the recent developments given by [31, 30, 32].

2 Preliminary

In this section, we give the problem formulation, the notations and the assumptions required for the convergence analysis.

2.1 Problem Formulation

Suppose that we are given n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ distributed from a probability distribution P on $\mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is an input space. We denote by Π the marginal distribution of P on $\mathcal{X}$. We are given M reproducing kernel Hilbert spaces (RKHS) $\{\mathcal{H}_m\}_{m=1}^M$ each of which is associated with a kernel k_m. We consider a mixed-norm type regularization with respect to an arbitrary given norm $\|\cdot\|_\psi$, that is, the regularization is given by the norm $\|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$ of the vector $(\|f_m\|_{\mathcal{H}_m})_{m=1}^M$ for $f_m \in \mathcal{H}_m$ $(m = 1, \ldots, M)$ *. For notational simplicity, we write $\|f\|_\psi = \|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$ for $f = \sum_{m=1}^M f_m$ $(f_m \in \mathcal{H}_m)$.

The general formulation of MKL that we consider in this article fits a function $f = \sum_{m=1}^M f_m$ $(f_m \in \mathcal{H}_m)$ to the data by solving the following optimization problem:

$$\hat{f} = \sum_{m=1}^M \hat{f}_m = \underset{f_m \in \mathcal{H}_m \ (m=1,\ldots,M)}{\arg\min} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{m=1}^M f_m(x_i) \right)^2 + \lambda_1^{(n)} \|f\|_\psi^2. \tag{1}$$

We call this "ψ-norm MKL". This formulation covers many practically used MKL methods (e.g., ℓ_p-MKL, elasticnet-MKL, variable sparsity kernel learning (see later for their definitions)), and is solvable by a finite dimensional optimization procedure due to the representer theorem [12]. In this article, we focus on the regression problem (the squared loss). However the discussion presented here can be generalized to Lipschitz continuous and strongly convex losses [6, 30].

Example 1: ℓ_p-MKL The first example of ψ-norm MKL is ℓ_p-MKL [14] that employs ℓ_p-norm for $1 \leq p \leq \infty$ as the regularizer: $\|f\|_\psi = \|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_{\ell_p} = (\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p)^{\frac{1}{p}}$. If p is strictly greater than 1 ($p > 1$), the solution of ℓ_p-MKL becomes dense. In particular, $p = 2$ corresponds to averaging candidate kernels with uniform weight [22]. It is reported that ℓ_p-MKL with p greater than 1, say $p = \frac{4}{3}$, often shows better performance than the original sparse ℓ_1-MKL [10].

*We assume that the mixed-norm $\|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$ satisfies the triangular inequality with respect to $(f_m)_{m=1}^M$, that is, $\|(\|f_m + f'_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi \leq \|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi + \|(\|f'_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$. To satisfy this condition, it is sufficient if the norm is monotone, i.e., $\|\boldsymbol{a}\|_\psi \leq \|\boldsymbol{a} + \boldsymbol{b}\|_\psi$ for all $\boldsymbol{a}, \boldsymbol{b} \geq \boldsymbol{0}$.

Example 2: Elasticnet-MKL The second example is elasticnet-MKL [26, 34] that employs mixture of ℓ_1 and ℓ_2 norms as the regularizer: $\|f\|_\psi = \tau\|f\|_{\ell_1} + (1-\tau)\|f\|_{\ell_2} = \tau\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m} + (1-\tau)(\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2)^{\frac{1}{2}}$ with $\tau \in [0,1]$. Elasticnet-MKL shares the same spirit with ℓ_p-MKL in a sense that it bridges sparse ℓ_1-regularization and dense ℓ_2-regularization. An efficient optimization method for elasticnet-MKL is proposed by [33].

Example 3: Variable Sparsity Kernel Learning Variable Sparsity Kernel Learning (VSKL) proposed by [1] divides the RKHSs into M' groups $\{\mathcal{H}_{j,k}\}_{k=1}^{M_j}$, $(j = 1, \ldots, M')$ and imposes a mixed norm regularization $\|f\|_\psi = \|f\|_{(p,q)} = \left\{\sum_{j=1}^{M'}(\sum_{k=1}^{M_j}\|f_{j,k}\|_{\mathcal{H}_{j,k}}^p)^{\frac{q}{p}}\right\}^{\frac{1}{q}}$ where $1 \leq p, q$, and $f_{j,k} \in \mathcal{H}_{j,k}$. An advantageous point of VSKL is that by adjusting the parameters p and q, various levels of sparsity can be introduced, that is, the parameters can control the level of sparsity *within* group and *between* groups. This point is beneficial especially for multi-modal tasks like object categorization.

2.2 Notations and Assumptions

Here, we prepare notations and assumptions that are used in the analysis. Let $\mathcal{H}^{\oplus M} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_M$. Throughout the article, we assume the following technical conditions (see also [3]).

Assumption 1. (Basic Assumptions)

(A1) *There exists* $f^* = (f_1^*, \ldots, f_M^*) \in \mathcal{H}^{\oplus M}$ *such that* $\mathrm{E}[Y|X] = f^*(X) = \sum_{m=1}^M f_m^*(X)$, *and the noise* $\epsilon := Y - f^*(X)$ *is bounded as* $|\epsilon| \leq L$.

(A2) *For each* $m = 1, \ldots, M$, $\mathcal{H}_m$ *is separable (with respect to the RKHS norm) and* $\sup_{X\in\mathcal{X}} |k_m(X,X)| < 1$.

The first assumption in (A1) ensures the model $\mathcal{H}^{\oplus M}$ is correctly specified, and the technical assumption $|\epsilon| \leq L$ allows ϵf to be Lipschitz continuous with respect to f. The noise boundedness can be relaxed to unbounded situation as in [24], but we don't pursue that direction for simplicity.

Let an integral operator $T_{k_m} : L_2(\Pi) \to L_2(\Pi)$ corresponding to a kernel function k_m be

$$T_{k_m} f = \int k_m(\cdot, x) f(x) \mathrm{d}\Pi(x).$$

It is known that this operator is compact, positive, and self-adjoint (see Theorem 4.27 of [28]). Thus it has at most countably many non-negative eigenvalues. We denote by $\mu_{\ell,m}$ be the ℓ-th largest eigenvalue (with possible multiplicity) of the integral operator T_{k_m}. Then we assume the following assumption on the decreasing rate of $\mu_{\ell,m}$.

Assumption 2. (Spectral Assumption) *There exist* $0 < s_m < 1$ *and* $0 < c$ *such that*

$$\text{(A3)} \qquad \mu_{\ell,m} \leq c\ell^{-\frac{1}{s_m}}, \quad (\forall \ell \geq 1,\ 1 \leq \forall m \leq M),$$

where $\{\mu_{\ell,m}\}_{\ell=1}^\infty$ *is the spectrum of the operator* T_{k_m} *corresponding to the kernel* k_m.

It was shown that the spectral assumption (A3) is equivalent to the classical covering number assumption [29]. Recall that the ϵ-covering number $N(\epsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi))$ with respect to $L_2(\Pi)$ is the minimal number of balls with radius ϵ needed to cover the unit ball $\mathcal{B}_{\mathcal{H}_m}$ in $\mathcal{H}_m$ [36]. If the spectral assumption (A3) holds, there exists a constant C that depends only on s and c such that $\log N(\varepsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi)) \leq C\varepsilon^{-2s_m}$, and the converse is also true (see [29, Theorem 15] and [28] for details). Therefore, if s_m is large, the RKHSs are regarded as "complex", and if s_m is small, the RKHSs are "simple".

An important class of RKHSs where s_m is known is Sobolev space. (A3) holds with $s_m = \frac{d}{2\alpha}$ for Sobolev space of α-times continuously differentiability on the Euclidean ball of $\mathbb{R}^d$ [11]. Moreover, for α-times continuously differentiable kernels on a closed Euclidean ball in $\mathbb{R}^d$, that holds for $s_m = \frac{d}{2\alpha}$ [28, Theorem 6.26]. According to Theorem 7.34 of [28], for Gaussian kernels with compact support distribution, that holds for arbitrary small $0 < s_m$. The covering number of Gaussian kernels with *unbounded* support distribution is also described in Theorem 7.34 of [28].

Let κ_M be defined as follows:

$$\kappa_M := \sup\left\{\kappa \geq 0 \;\middle|\; \kappa \leq \frac{\|\sum_{m=1}^M f_m\|_{L_2(\Pi)}^2}{\sum_{m=1}^M \|f_m\|_{L_2(\Pi)}^2},\ \forall f_m \in \mathcal{H}_m\ (m = 1, \ldots, M)\right\}. \tag{2}$$

κ_M represents the correlation of RKHSs. We assume all RKHSs are not completely correlated to each other.

Assumption 3. (Incoherence Assumption) *κ_M is strictly bounded from below; there exists a constant $C_0 > 0$ such that*

(A4) $$0 < C_0^{-1} < \kappa_M.$$

This condition is motivated by the *incoherence condition* [18, 21] considered in sparse MKL settings. This ensures the uniqueness of the decomposition $f^* = \sum_{m=1}^M f_m^*$ of the ground truth. [3] also assumed this condition to show the consistency of ℓ_1-MKL.

Finally we give a technical assumption with respect to ∞-norm.

Assumption 4. (Embedded Assumption) *Under the Spectral Assumption, there exists a constant $C_1 > 0$ such that*

(A5) $$\|f_m\|_\infty \leq C_1 \|f_m\|_{\mathcal{H}_m}^{1-s_m} \|f_m\|_{L_2(\Pi)}^{s_m}.$$

The condition (A5) is common and practical. There is a clear characterization of the condition (A5) in terms of *real interpolation of spaces*. One can find detailed and formal discussions of interpolations in [29], and Proposition 2.10 of [7] gives the necessary and sufficient condition for the assumption (A5).

3 Convergence Rate Analysis of ψ-norm MKL

Here we present the learning rate of ψ-norm MKL derived by [31, 30]. We suppose that the number of kernels M can increase along with the number of samples n.

Now we define $\eta(t) := \eta_n(t) = \max(1, \sqrt{t}, t/\sqrt{n})$ for $t > 0$, and, for given positive reals $\{r_m\}_{m=1}^M$ and given n, we define $\alpha_1, \alpha_2, \beta_1, \beta_2$ as follows:

$$\alpha_1 := \alpha_1(\{r_m\}) = 3\left(\sum_{m=1}^M \frac{r_m^{-2s_m}}{n}\right)^{\frac{1}{2}}, \quad \alpha_2 := \alpha_2(\{r_m\}) = 3\left\|\left(\frac{s_m r_m^{1-s_m}}{\sqrt{n}}\right)_{m=1}^M\right\|_{\psi^*},$$

$$\beta_1 := \beta_1(\{r_m\}) = 3\left(\sum_{m=1}^M \frac{r_m^{-\frac{2s_m(3-s_m)}{1+s_m}}}{n^{\frac{2}{1+s_m}}}\right)^{\frac{1}{2}}, \quad \beta_2 := \beta_2(\{r_m\}) = 3\left\|\left(\frac{\frac{(1-s_m)^2}{1+s_m}}{\frac{s_m r_m}{n^{\frac{1}{1+s_m}}}}\right)_{m=1}^M\right\|_{\psi^*}, \quad (3)$$

(note that $\alpha_1, \alpha_2, \beta_1, \beta_2$ implicitly depends on the reals $\{r_m\}_{m=1}^M$). Then the following theorem gives the general form of the learning rate of ψ-norm MKL.

Theorem 1 ([31, 30]). *Suppose Assumptions 1-4 are satisfied. Let $\{r_m\}_{m=1}^M$ be arbitrary positive reals that can depend on n, and assume $\lambda_1^{(n)} = \left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2$. Then for all n and t' that satisfy $\frac{\log(M)}{\sqrt{n}} \leq 1$ and $\frac{4\phi\sqrt{n}}{\kappa_M}\max\{\alpha_1^2, \beta_1^2, \frac{M\log(M)}{n}\}\eta(t') \leq \frac{1}{12}$ and for all $t \geq 1$, we have*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{24\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2 + \beta_1^2 + \frac{M\log(M)}{n}\right) + 4\left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2. \quad (4)$$

with probability $1 - \exp(-t) - \exp(-t')$.

The statement of Theorem 1 itself is complicated. Thus we will show later concrete learning rates on some examples such as ℓ_p-MKL. The convergence rate (4) depends on the positive reals $\{r_m\}_{m=1}^M$, but the choice of $\{r_m\}_{m=1}^M$ are arbitrary. Thus by minimizing the right hand side of Eq. (4), we obtain tight convergence bound as follows:

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(\min_{\substack{\{r_m\}_{m=1}^M: \\ r_m > 0}}\left\{\alpha_1^2 + \beta_1^2 + \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2 + \frac{M\log(M)}{n}\right\}\right). \quad (5)$$

There is a trade-off between the first two terms $(a) := \alpha_1^2 + \beta_1^2$ and the third term $(b) := \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2$, that is, if we take $\{r_m\}_m$ large, then the term (a) becomes small and the term (b) becomes large, on the other hand, if we take $\{r_m\}_m$ small, then it results in large (a) and small (b). Therefore we need to balance the two terms (a) and (b) to obtain the minimum in Eq. (5).

We discuss the obtained learning rate in two situations, (i) *homogeneous complexity* situation, and (ii) *inhomogeneous complexity* situation:

(i) (homogeneous) All s_ms are same: there exists $0 < s < 1$ such that $s_m = s$ $(\forall m)$ (Sec.3.1).

(ii) (inhomogeneous) All s_ms are *not* same: there exist m, m' such that $s_m \neq s_{m'}$ (Sec.3.2).

3.1 Analysis on Homogeneous Settings

Here we assume all s_ms are same, say $s_m = s$ for all m (homogeneous setting). If we further restrict the situation as all r_ms are same ($r_m = r$ $(\forall m)$ for some r), then the minimization in Eq. (5) can be easily carried out as in the following corollary. Let $\mathbf{1}$ be the M-dimensional vector each element of which is 1: $\mathbf{1} := (1, \ldots, 1)^\top \in \mathbb{R}^M$, and $\|\cdot\|_{\psi^*}$ be the dual norm of the ψ-norm[†].

Corollary 2. *When $s_m = s$ $(\forall m)$ with some $0 < s < 1$ and $n \geq (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi/M)^{\frac{4s}{1-s}}$, the bound* (5) *indicates that*

$$\|\hat{f} - f^*\|^2_{L_2(\Pi)} = \mathcal{O}_p\left(M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right). \tag{6}$$

Corollary 2 is derived by assuming $r_m = r$ $(\forall m)$, which might make the bound loose. However, when the norm $\|\cdot\|_\psi$ is *isotropic* (whose definition will appear later), that restriction ($r_m = r$ $(\forall m)$) does not make the bound loose, that is, the upper bound obtained in Corollary 2 is tight and achieves the minimax optimal rate (the minimax optimal rate is the one that cannot be improved by any estimator). In the following, we investigate the general result of Corollary 2 through some important examples.

Convergence Rate of ℓ_p-MKL Here we derive the convergence rate of ℓ_p-MKL ($1 \leq p \leq \infty$) where $\|f\|_\psi = \sum_{m=1}^M (\|f_m\|^p_{\mathcal{H}_m})^{\frac{1}{p}}$ (for $p = \infty$, it is defined as $\max_m \|f_m\|_{\mathcal{H}_m}$). It is well known that the dual norm of ℓ_p-norm is given as ℓ_q-norm where q is the real satisfying $\frac{1}{p} + \frac{1}{q} = 1$. For notational simplicity, let $R_p := \left(\sum_{m=1}^M \|f^*_m\|^p_{\mathcal{H}_m}\right)^{\frac{1}{p}}$. Then substituting $\|f^*\|_\psi = R_p$ and $\|\mathbf{1}\|_{\psi^*} = \|\mathbf{1}\|_{\ell_q} = M^{\frac{1}{q}} = M^{1-\frac{1}{p}}$ into the bound (6), the learning rate of ℓ_p-MKL is given as

$$\|\hat{f} - f^*\|^2_{L_2(\Pi)} = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}} R_p^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right). \tag{7}$$

If we further assume n is sufficiently large so that $n \geq M^{\frac{2}{p}} R_p^{-2} (\log M)^{\frac{1+s}{s}}$, the leading term is the first term, and thus we have

$$\|\hat{f} - f^*\|^2_{L_2(\Pi)} = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}} R_p^{\frac{2s}{1+s}}\right). \tag{8}$$

Note that as the complexity s of RKHSs becomes small the convergence rate becomes fast. It is known that $n^{-\frac{1}{1+s}}$ is the minimax optimal learning rate for single kernel learning. The derived rate of ℓ_p-MKL is obtained by multiplying a coefficient depending on M and R_p to the optimal rate of single kernel learning. To investigate the dependency of R_p to the learning rate, let us consider two extreme settings, i.e., sparse setting $(\|f^*_m\|_{\mathcal{H}_m})_{m=1}^M = (1, 0, \ldots, 0)$ and dense setting $(\|f^*_m\|_{\mathcal{H}_m})_{m=1}^M = (1, \ldots, 1)$ as in [15].

[†]The dual of the norm $\|\cdot\|_\psi$ is defined as $\|\boldsymbol{b}\|_{\psi^*} := \sup_{\boldsymbol{a}}\{\boldsymbol{b}^\top \boldsymbol{a} \mid \|\boldsymbol{a}\|_\psi \leq 1\}$.

- $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = (1, 0, \ldots, 0)$: $R_p = 1$ for all p. Therefore the convergence rate $n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}}$ is fast for small p and the minimum is achieved at $p = 1$. This means that ℓ_1 regularization is preferred for sparse truth.

- $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = (1, \ldots, 1)$: $R_p = M^{\frac{1}{p}}$, thus the convergence rate is $Mn^{-\frac{1}{1+s}}$ for all p. Interestingly for dense ground truth, there is no dependency of the convergence rate on the parameter p (later we will show that this is not the case in inhomogeneous settings (Sec.3.2)). That is, the convergence rate is M times the optimal learning rate of single kernel learning ($n^{-\frac{1}{1+s}}$) for all p. This means that for the dense settings, the complexity of solving MKL problem is equivalent to that of solving M single kernel learning problems.

Comparison with Existing Bounds Here we present a comparison of the bound for ℓ_p-MKL shown above with other existing bounds. Let $\mathcal{H}_{\ell_p}(R)$ be the ℓ_p-mixed norm ball with radius R: $\mathcal{H}_{\ell_p}(R) := \{f = \sum_{m=1}^M f_m \mid (\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p)^{\frac{1}{p}} \leq R\}$. [10, 16, 15] gave "global" type bounds for ℓ_p-MKL as

$$R(f) \leq \widehat{R}(f) + C\frac{M^{1-\frac{1}{p}} \vee \sqrt{\log(M)}}{\sqrt{n}} R \quad \text{for all } f \in \mathcal{H}_{\ell_p}(R), \tag{9}$$

where $R(f)$ and $\widehat{R}(f)$ is the population risk and the empirical risk. First observation is that the bounds by [10] and [15] are restricted to the situation $1 \leq p \leq 2$. On the other hand, the presented analysis and that of [16] covers all $p \geq 1$. Second, since the bound (8) is specialized to the regularized risk minimizer $\hat{f}$ defined at Eq. (1) while the bound (9) is applicable to all $f \in \mathcal{H}_{\ell_p}(R)$, the bound (8) is sharper than theirs for sufficiently large n. To see this, suppose $n \geq M^{\frac{2}{p}} R_p^{-2}$, then we have $n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}} \leq n^{-\frac{1}{2}} M^{1-\frac{1}{p}}$. Moreover we should note that s can be large as long as Spectral Assumption (A3) is satisfied. Thus the bound (9) is formally recovered by our analysis by approaching s to 1.

Recently [13] gave a tighter convergence rate utilizing the localization technique as $\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\Big(\min_{p' \geq p} \Big\{\frac{p'}{p'-1} n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p'(1+s)}} R_{p'}^{\frac{2s}{1+s}}\Big\}\Big)$. Comparing the presented bound (8) with their result, there are not $\min_{p' \geq p}$ and $\frac{p'}{p'-1}$ in the bound (8) (if there is not the term $\frac{p'}{p'-1}$, then the minimum of $\min_{p' \geq p}$ is attained at $p' = p$, thus the bound (8) is tighter).

Convergence Rate of Elasticnet-MKL Elasticnet-MKL employs a mixture of ℓ_1 and ℓ_2 norm as the regularizer: $\|f\|_\psi = \tau\|f\|_{\ell_1} + (1-\tau)\|f\|_{\ell_2}$ where $\tau \in [0, 1]$. Then its dual norm is given by $\|b\|_{\psi^*} = \min_{a \in \mathbb{R}^M} \Big\{\max\Big(\frac{\|a\|_{\ell_\infty}}{\tau}, \frac{\|a-b\|_{\ell_2}}{1-\tau}\Big)\Big\}$. Therefore by a simple calculation, we have $\|\mathbf{1}\|_{\psi^*} = \frac{\sqrt{M}}{1-\tau+\tau\sqrt{M}}$. Hence Eq. (6) gives the convergence rate of elasticnet-MKL as

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} \frac{M^{1-\frac{s}{1+s}}}{(1-\tau+\tau\sqrt{M})^{\frac{2s}{1+s}}} (\tau\|f^*\|_{\ell_1} + (1-\tau)\|f^*\|_{\ell_2})^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right).$$

Note that, when $\tau = 0$ or $\tau = 1$, this rate is identical to that of ℓ_2-MKL or ℓ_1-MKL obtained in Eq. (7) respectively.

3.1.1 Minimax Lower Bound

It can be shown that the presented learning rate (6) achieves the minimax-learning rate on the ψ-norm ball

$$\mathcal{H}_\psi(R) := \left\{ f = \textstyle\sum_{m=1}^M f_m \mid \|f\|_\psi \le R \right\},$$

when the norm is *isotropic*. We say the ψ-norm $\|\cdot\|_\psi$ is isotropic when there exits a universal constant $\bar{c}$ such that

$$\bar{c}M = \bar{c}\|\mathbf{1}\|_{\ell_1} \ge \|\mathbf{1}\|_{\psi^*}\|\mathbf{1}\|_\psi, \qquad \|\boldsymbol{b}\|_\psi \le \|\boldsymbol{b}'\|_\psi \ \ (\text{if } 0 \le b_m \le b'_m \ (\forall m)), \tag{10}$$

(note that the inverse inequality $M \le \|\mathbf{1}\|_{\psi^*}\|\mathbf{1}\|_\psi$ of the first condition always holds by the definition of the dual norm). Practically used regularizations usually satisfy this isotropic property. In fact, ℓ_p-MKL, elasticnet-MKL and VSKL satisfy the isotropic property with $\bar{c} = 1$.

To derive the minimax learning rate, we consider a simpler situation. First we assume that each RKHS is same as others. That is, the input vector is decomposed into M components like $x = (x^{(1)}, \ldots, x^{(M)})$ where $\{x^{(m)}\}_{m=1}^M$ are M i.i.d. copies of a random variable $\tilde{X}$, and $\mathcal{H}_m = \{f_m \mid f_m(x) = f_m(x^{(1)}, \ldots, x^{(M)}) = \tilde{f}_m(x^{(m)}),\ \tilde{f}_m \in \widetilde{\mathcal{H}}\}$ where $\widetilde{\mathcal{H}}$ is an RKHS shared by all $\mathcal{H}_m$. Thus $f \in \mathcal{H}^{\oplus M}$ is decomposed as $f(x) = f(x^{(1)}, \ldots, x^{(M)}) = \sum_{m=1}^M \tilde{f}_m(x^{(m)})$ where each $\tilde{f}_m$ is a member of the common RKHS $\widetilde{\mathcal{H}}$. We denote by $\tilde{k}$ the kernel associated with the RKHS $\widetilde{\mathcal{H}}$.

In addition to the condition about the upper bound of spectrum (Spectral Assumption (A3)), we assume that the spectrum of all the RKHSs $\{\mathcal{H}_m\}_{m=1}^M$ have the same lower bound of polynomial rate.

Assumption 5. (Strong Spectral Assumption) *There exist* $0 < s < 1$ *and* $0 < c, c'$ *such that*

$$c'\ell^{-\frac{1}{s}} \le \tilde{\mu}_\ell \le c\ell^{-\frac{1}{s}}, \quad (1 \le \forall \ell), \tag{A6}$$

where $\{\tilde{\mu}_\ell\}_{\ell=1}^\infty$ *is the spectrum of the integral operator* $T_{\tilde{k}}$ *corresponding to the kernel* $\tilde{k}$. *In particular, the spectrum of* T_{k_m} *also satisfies* $\mu_{\ell,m} \sim \ell^{-\frac{1}{s}}$ $(\forall \ell, m)$.

Without loss of generality, we may assume that $\mathrm{E}[f(\tilde{X})] = 0 \quad (\forall f \in \widetilde{\mathcal{H}})$. Since each f_m receives i.i.d. copy of $\tilde{X}$, $\mathcal{H}_m$s are orthogonal to each other:

$$\mathrm{E}[f_m(X)f_{m'}(X)] = \mathrm{E}[\tilde{f}_m(X^{(m)})\tilde{f}_{m'}(X^{(m')})] = 0 \quad (\forall f_m \in \mathcal{H}_m,\ \forall f_{m'} \in \mathcal{H}_{m'},\ \forall m \ne m').$$

We also assume that the noise $\{\epsilon_i\}_{i=1}^n$ is an i.i.d. normal sequence with standard deviation $\sigma > 0$.

Under the assumptions described above, we have the following minimax $L_2(\Pi)$-error.

Theorem 3 ([31, 30]). *Suppose* $R > 0$ *is given and* $n > \frac{\bar{c}^2 M^2}{R^2\|\mathbf{1}\|_{\psi^*}^2}$ *is satisfied. Then the minimax-learning rate on* $\mathcal{H}_\psi(R)$ *for isotropic norm* $\|\cdot\|_\psi$ *is lower bounded as*

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_\psi(R)} \mathrm{E}\left[\|\hat{f} - f^*\|_{L_2(\Pi)}^2\right] \ge CM^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*} R)^{\frac{2s}{1+s}}, \tag{11}$$

where inf *is taken over all measurable functions of* n *samples* $\{(x_i, y_i)\}_{i=1}^n$.

One can see that the convergence rate derived in Eq. (8) achieves the minimax rate on the ψ-norm ball (Theorem 3) up to $\frac{M\log(M)}{n}$ that is negligible when the number of samples is large. Indeed if

$$n \geq \frac{M^2 \log(M)^{\frac{1+s}{s}}}{\|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2}, \tag{12}$$

then the first term in Eq. (8) dominates the second term $\frac{M\log(M)}{n}$ and the upper bound coincides with the minimax optima rate. Note that the condition (12) for the sample size n is equivalent to the condition for n assumed in Theorem 3 up to factors of $\log(M)^{\frac{1+s}{s}}$ and a constant.

By the definition of the dual norm, one can check that the norm that minimizes this bound (6) is the ℓ_1-norm. Moreover, if ψ-norm is isotropic, the bound is tight and can not be improved as shown in Theorem 3. Therefore, ℓ_1-norm is the optimal regularization among all isotropic norms in homogeneous settings. However if ψ-norm is not isotropic, the bound is no longer tight. That means non-isotropic norms can outperform isotropic norms if the non-isotropic norm is appropriately chosen. In particular, ℓ_1-norm can be outperformed by some non-isotropic norm for a particular choice of f^*. In Section 4, we introduce an adaptive method that utilizes a non-isotropic norm regularization specifically tailored to the truth f^*.

3.2 Analysis on Inhomogeneous Settings

In the previous section (analysis on homogeneous settings), we have not seen any theoretical justification supporting the fact that dense MKL methods like $\ell_{\frac{4}{3}}$-MKL can outperform the sparse ℓ_1-MKL [10]. However, it can be shown that dense type regularizations can outperform the sparse regularization in inhomogeneous settings (there exists m, m' such that $s_m \neq s_{m'}$). For simplicity, we focus on ℓ_p-MKL, and discuss the relation between the learning rate and the norm parameter p.

Let us consider an extreme situation where $s_1 = s$ for some $0 < s < 1$ and $s_m = 0$ $(m > 1)$[‡]. In this situation, we have

$$\alpha_1 = 3\left(\frac{r_1^{-2s}+M-1}{n}\right)^{\frac{1}{2}},\ \alpha_2 = 3\frac{sr_1^{1-s}}{\sqrt{n}},\ \beta_1 = 3\left(\frac{r_1^{-\frac{2s(3-s)}{1+s}}+M-1}{n^{\frac{2}{1+s}}}\right)^{\frac{1}{2}},\ \beta_2 = 3\frac{sr_1^{\frac{(1-s)^2}{1+s}}}{n^{\frac{1}{1+s}}}.$$

for all p. Note that these α_1, α_2, β_1 and β_2 have no dependency on p. Therefore the learning bound (5) is smallest when $p = \infty$ because $\|f^*\|_{\ell_\infty} \leq \|f^*\|_{\ell_p}$ for all $1 \leq p < \infty$. In particular, when $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = \mathbf{1}$, we have $\|f^*\|_{\ell_1} = M\|f^*\|_{\ell_\infty}$ and thus obviously the learning rate of ℓ_∞-MKL given by Eq. (5) is faster than that of ℓ_1-MKL. In fact, through a bit cumbersome calculation, one can check that ℓ_∞-MKL can be $M^{\frac{2s}{1+s}}$ times faster than ℓ_1-MKL in a worst case. This indicates that, when the complexities of RKHSs are inhomogeneous, the generalization abilities of *dense* type regularizations (e.g., ℓ_∞-MKL) can be better than the *sparse* type regularization (ℓ_1-MKL). In real settings, it is likely that

[‡]In our assumption s_m should be greater than 0. However we formally put $s_m = 0$ $(m > 1)$ for simplicity of discussion. For rigorous discussion, one might consider arbitrary small $s_m \ll s$.

one uses various types of kernels and the complexities of RKHSs become inhomogeneous. As mentioned above, it has been often reported that ℓ_1-MKL is outperformed by dense type MKL such as $\ell_{\frac{4}{3}}$-MKL in numerical experiments [10]. The theoretical analysis explains well this experimental results.

4 Adaptively weighted estimator

[32] proposed a two-stage method that adaptively make use of a non-isotropic norm regularization. The estimating procedure is as follows: In the first stage, we prepare a rough estimator $\tilde{f} = \sum_{m=1}^{M} \tilde{f}_m$, then, in the second stage, we compute the ψ-norm MKL estimator where, as the regularization term, we employ the following norm based on the rough estimator $\tilde{f}$:

$$\|f\|_{\psi,\gamma} := \|(\|f_m\|_{\mathcal{H}_m}/\|\tilde{f}_m\|_{\mathcal{H}_m}^{\gamma})_{m=1}^{M}\|_{\psi}.$$

This estimator is called an adaptively weighted estimator. Note that, when $\gamma = 0$, the adaptively weighted estimator is just the normal ψ-norm MKL. In general, the norm $\|f\|_{\psi,\gamma}$ is not isotropic for $\gamma > 0$ even if $\|\cdot\|_{\psi}$ is isotropic. Suppose the rough estimator $\tilde{f}$ well approximate the true function f^*, then the adaptively weighted estimator imposes a large penalty on the components where f_m^* is small and imposes a small penalty on the components of large f_m^*. Intuitively the adaptive estimator amplifies important components and diminishes unimportant components. The parameter γ controls the strength of the adaptivity. This kind of idea is already proposed in a linear regression model as an *adaptive lasso* [40]. The adaptively weighted estimator can be seen as its MKL version.

To see the effectiveness of the method, we give an informal discussion on an extreme situation where $\tilde{f}_m = f_m^*$ for all m, $f_m^* = \tilde{f}_m = 0$ for $m = 2, \ldots, M$, and $\|f_1^*\|_{\mathcal{H}_1} = 1$. For simplicity, we assume $\gamma < 1$ and use a convention $\|f_m^*\|_{\mathcal{H}_m}/\|f_m^*\|_{\mathcal{H}_m}^{\gamma} = \|f_m^*\|_{\mathcal{H}_m}^{1-\gamma} = 0$ for $m = 2, \ldots, M$. In this situation, letting $\|\cdot\|_{\psi^*,\gamma}$ be the dual norm of $\|\cdot\|_{\psi,\gamma}$, we have $\|\boldsymbol{a}\|_{\psi^*,\gamma} = \|(\boldsymbol{a}_1, 0, \ldots, 0)\|_{\psi^*,\gamma}$. Hence we can check that using the bound (5) the adaptively weighted estimator $\check{f}$ yields the following learning rate:

$$\|\check{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_{\psi})^{\frac{2s}{1+s}}\right),$$

for sufficiently large n. This learning rate is $M^{1-\frac{2s}{1+s}}$ times faster than the bound (6). This (informal) discussion indicates that, if f^* is well approximated by $\tilde{f}$, the adaptively weighted estimator yields a better performance than the non-adaptive one.

5 Numerical Experiments

5.1 Comparison between Homogeneous and Inhomogeneous Settings

Here we investigate numerically how the inhomogeneity of the complexities affects the performances using synthetic data. In particular, we numerically compare two situations:

(a) all complexities of RKHSs are same (homogeneous situation) and (b) one RKHS is complex and other RKHSs are evenly simple (inhomogeneous situation)[§].

The experimental settings are as follows. The input random variable is 20 dimensional vector $x = (x^{(1)}, \ldots, x^{(20)})$ where each element $x^{(m)}$ is independently identically distributed from the uniform distribution on $[0, 1]$: $x^{(m)} \sim \mathrm{Unif}([0, 1])$ $(m = 1, \ldots, 20)$. For each coordinate $m = 1, \ldots, 20$, we put one Gaussian RKHS $\mathcal{H}_m$ with a Gaussian width σ_m: the number of kernels is 20 $(M = 20)$ and

$$k_m(x, x') = \exp\left(-\frac{(x^{(m)} - x'^{(m)})^2}{2\sigma_m^2}\right) \quad (m = 1, \ldots, 20),$$

for $x = (x^{(1)}, \ldots, x^{(20)})$ and $x' = (x'^{(1)}, \ldots, x'^{(20)})$. To generate the ground truth f^*, we randomly generated 5 center points $\mu_{i,m}$ $(i = 1, \ldots, 5)$ for each coordinate $m = 1, \ldots, 20$ where $\mu_{i,m}$ is independently generated by the uniform distribution on $[0, 1]$. Then we obtain the following form of the true function:

$$f^*(x) = \sum_{m=1}^{20} f_m^*(x), \quad \text{where} \quad f_m^*(x) = \sum_{i=1}^{5} \alpha_{i,m} \exp\left(-\frac{(x^{(m)} - \mu_{i,m})^2}{2\sigma_m^2}\right) \in \mathcal{H}_m,$$

for $x = (x_1, \ldots, x_m)$. Each coefficient $\alpha_{i,m}$ is independently identically distributed from the standard normal distribution. The output y is contaminated by a noise ϵ where the noise ϵ is distributed from the Gaussian distribution with mean 0 and standard deviation 0.1: $y = f_m^*(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$.

We generated 200 realizations $\{(x_i, y_i)\}_{i=1}^n$ $(n = 200)$, and estimated f^* using ℓ_p-MKL with $p = 1, 1.1, 1.2, \ldots, 3$ [¶]. The estimator is computed with various regularization parameters $\lambda_1^{(n)}$. The generalization error $\|\hat{f} - f^*\|_{L_2(\Pi)}^2$ was numerically calculated. We repeated the experiments for 100 times, averaged the generalization errors over 100 repetitions for each p and each regularization parameter, and obtained the optimal average generalization error among all regularization parameters for each p. The true function was randomly generated for each repetition. We investigated the generalization errors in the following homogeneous and inhomogeneous settings:

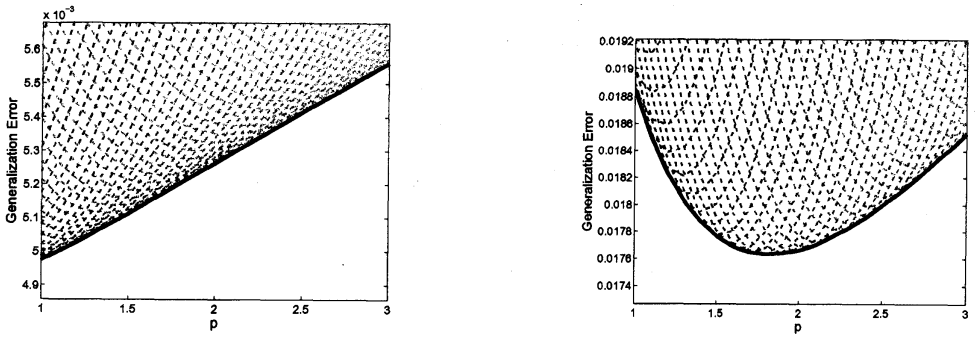
1. (homogeneous) $\sigma_m = 0.5$ for $m = 1, \ldots, 20$.
2. (inhomogeneous) $\sigma_1 = 0.01$ and $\sigma_m = 0.5$ for $m = 2, \ldots, 20$.

The difference between the above homogeneous and inhomogeneous settings is the value of σ_1; whether $\sigma_1 = 0.5$ or $\sigma_1 = 0.01$.

Figure 1 shows the average generalization errors in (a) the homogeneous setting, and (b) the inhomogeneous setting. Each broken line corresponds to one regularization parameter. The bold solid line shows the best (average) generalization error among all the regularization parameters. We can see that in the homogeneous setting ℓ_1-regularization

[§]More detailed descriptions about the experiment can be found in [30].

[¶]We included a bias term in this experiment, that is, we fitted $\hat{f}(x)+b$ to the data: $\min_{f_m, b} \frac{1}{n}\sum_{i=1}^n (y_i - \sum_{m=1}^M f_m(x_i) - b)^2 + \lambda_1^{(n)} \|f\|_{\ell_p}^2$.



(a) Homogeneous Setting

(b) Inhomogeneous Setting

Figure 1: The expected generalization error $\mathrm{E}[\|\hat{f} - f^*\|^2_{L_2(\Pi)}]$ against the parameter p for ℓ_p-MKL. Each broken line corresponds to one regularization parameter. The bold solid line shows the best generalization error among all the regularization parameters.

shows the best performance, on the other hand, in the inhomogeneous setting the best performance is achieved at $p > 1$. This experimental result matches the theoretical investigations.

5.2 Evaluation of Adaptively Weighted Estimator

In this section, we present a numerical experiment that demonstrates the effectiveness of the adaptively weighted estimator presented in Section 4[‖]. 13 datasets included in the IDA benchmark repository were used. All of them are binary classification tasks. Since the analyses in previous sections are about regression problems where the squared loss is employed, that can not be applied directly to binary classifications. However, there are tight relations between properties of classification and regression. Thus a performance analysis in regression problems gives the same qualitative evaluation also for classification tasks. The candidate kernels were Gaussian kernels with 10 different bandwidths (0.5 1 2 5 7 10 12 15 17 20) applied on jointly all the variables, Gaussian kernels with 5 different bandwidths (1 5 10 15 20) applied on individual variables and polynomial kernels of degree 1 to 3 applied on jointly all the variables. The total number of candidate kernels is $5 \times d + 13$, where d is the number of variables.

As the rough estimator $\tilde{f}$, we employed the ℓ_2-MKL estimator where the logistic loss is used. Then we computed the adaptively weighted estimator for ℓ_p-norm regularization with $p = (1.1, 4/3, 1.5, 2)$ and $\gamma = 0, 1, 2$. We repeated the experiments 20 times on different training-test sample combinations, and averaged the classification accuracies. We have three free parameters: the regularization constants $\lambda_1^{(n)}$ for the rough estimator and the second stage estimator and the parameter p for the second stage estimator. We chose the parameters that achieves the best averaged classification accuracy. Table 5.2 shows the averaged classification accuracy (%) for each γ and each datasets. Here again note that $\gamma = 0$ corresponds to the naive ℓ_p-MKL. We can see that the adaptively weighted

[‖]The experiment is originally presented in [32]

Table 1: The averaged classification accuracy % over 20 independent repetition. The best method in terms of the averaged accuracy is indicated by boldface.

Data	γ		
	0	1	2
banana	89.5	**89.5**	89.5
breast-cancer	74.2	**74.4**	74.4
diabetis	76.8	76.9	**77.0**
flare-solar	67.5	**67.7**	67.5
german	77.1	77.2	**77.3**
heart	84.4	**84.4**	84.2
image	97.2	97.5	**97.6**

Data	γ		
	0	1	2
ringnorm	97.4	**97.6**	97.5
splice	94.4	**94.9**	94.9
thyroid	95.9	96.1	**96.1**
titanic	77.9	78.0	**78.1**
twonorm	**97.6**	97.5	97.5
waveform	**90.0**	89.9	89.7

estimator ($\gamma = 1, 2$) shows favorable performances against the naive approach ($\gamma = 0$). This result supports the effectiveness of our proposed adaptive estimator.

Finally we would like to note that the incoherence assumption (A4) is not satisfied in this experiment. However the numerical experiment shows that, even if the assumption is not satisfied, the adaptively weighted estimator can give a favorable performance.

6 Conclusion

We reviewed the recently developed convergence analysis of MKL [31, 30] and the adaptively weighted estimator [32]. The unified framework gives the learning rate of MKL with arbitrary mixed-norm-type regularization. We have seen that the convergence rate of ℓ_p-MKL obtained in homogeneous settings is tighter than existing results. Moreover, the derived learning rate is minimax optimal. Furthermore, we observed that the bound well explains the favorable experimental results for dense type MKL by considering the inhomogeneous settings. We also presented the adaptively weighted estimator and observed its effectiveness through numerical experiments and an informal theoretical discussion.

Acknowledgement We would like to thank Marius Kloft, Gilles Blanchard, Ryota Tomioka and Masashi Sugiyama for suggestive discussions. This work was partially supported by MEXT Kakenhi 22700289 and the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

References

[1] J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. S. Nath, and S. Raman. Variable sparsity kernel learning. *Journal of Machine Learning Research*, 12:565–592, 2011.

[2] A. Argyriou, R. Hauser, C. A. Micchelli, and M. Pontil. A DC-programming algorithm for kernel selection. In *the 23st ICML*, pages 41–48, 2006.

[3] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[4] F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems 21*, pages 105–112, 2009.

[5] F. R. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *the 21st ICML*, pages 41–48, 2004.

[6] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33:1487–1537, 2005.

[7] C. Bennett and R. Sharpley. *Interpolation of Operators.* Academic Press, Boston, 1988.

[8] C. Cortes, M. Mohri, and A. Rostamizadeh. L_2 regularization for learning kernels. In *UAI 2009*, 2009.

[9] C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In *Advances in Neural Information Processing Systems 22*, pages 396–404, 2009.

[10] C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *the 27th ICML*, pages 247–254, 2010.

[11] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators.* Cambridge University Press, Cambridge, 1996.

[12] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[13] M. Kloft and G. Blanchard. On the convergence rate of ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 13:2465–2501, 2012.

[14] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate ℓ_p-norm multiple kernel learning. In *Advances in Neural Information Processing Systems 22*, pages 997–1005, 2009.

[15] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. l_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

[16] M. Kloft, U. Rückert, and P. L. Bartlett. A unifying view of multiple kernel learning. In *ECML/PKDD*, 2010.

[17] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

[18] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *COLT*, pages 229–238, 2008.

[19] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

[20] G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[21] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

[22] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

[23] C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, 2005.

[24] G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical report, 2010. arXiv:1008.3654.

[25] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[26] J. Shawe-Taylor. Kernel learning for novelty detection. In *NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, Whistler, 2008.

[27] N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *COLT*, pages 169–183, 2006.

[28] I. Steinwart. *Support Vector Machines.* Springer, 2008.

[29] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *COLT 2009*, 2009.

[30] T. Suzuki. Fast learning rate of non-sparse multiple kernel learning and optimal regularization strategies, 2011. arXiv:1111.3781.

[31] T. Suzuki. Unifying framework for fast learning rate of non-sparse multiple kernel learning. In *Advances in Neural Information Processing Systems 24*, pages 1575–1583, 2011.

[32] T. Suzuki. Improvement of multiple kernel learning using adaptively weighted regularization. *JSIAM Letters*, page to appear, 2013.

[33] T. Suzuki and R. Tomioka. Spicymkl: A fast algorithm for multiple kernel learning with thousands of kernels. *Machine Learning*, 85(1):77–108, 2011.

[34] R. Tomioka and T. Suzuki. Sparsity-accuracy trade-off in MKL. In *NIPS 2009 Workshop: Understanding Multiple Kernel Learning Methods*, Whistler, 2009.

[35] S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[36] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York, 1996.

[37] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *the 26th ICML*, pages 1065–1072, 2009.

[38] Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In *COLT*, 2009.

[39] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[40] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.